

REDECI: colaboração e produção científica em ciência da informação no Brasil

Fernando Silva Parreiras, Antonio Braz de Oliveira e Silva,
Renato Fabiano Matheus, Wladimir Cardoso Brandão

Apresenta a metodologia e os resultados preliminares da análise da co-autoria na área de ciência da informação no Brasil, através do uso da técnica conhecida como análise de rede social - ARS - , como parte do projeto *RedeCI*. Os resultados indicam a concentração de artigos com autoria única e de autores transientes. O número de autores com apenas uma contribuição é significativo e o resultado segue a Lei de Lotka. Os diferentes índices de centralidade na rede têm baixo índice de correlação entre si. Futuros desenvolvimentos permitirão uma boa compreensão dos colégios invisíveis existentes na área da ciência da informação.

Palavras-chave: Análise de redes sociais; Ciência da informação; Bibliometria; Cienciometria; Webometria.

Recebido em 27.03.2006

Aceito em 22.05.2006

1. Introdução

O conhecimento científico é uma questão essencial para a ciência da informação desde sua origem (BUSH, 1945). Tal fato coloca a recuperação da informação (SARACEVIC, 1996) e os estudos bibliométricos no centro das pesquisas na área (BROOKES, 1980). Dentre os estudos bibliométricos, estão os de co-autoria, que buscam entender a produção científica por meio da análise da colaboração entre pesquisadores. A análise de redes sociais - ARS - , por seu turno, oferece ferramentas capazes de analisar redes de contatos entre pessoas (WASSERMAN; FAUST, 1999), inclusive redes de colaboração.

O objetivo deste artigo é apresentar a metodologia e os resultados preliminares de uma pesquisa das redes de colaborações entre pesquisadores da área da ciência da informação no Brasil, doravante denominada *RedeCI*, enfatizando os métodos de ARS, sem se restringir a eles, e indicando como se deu a criação da base de dados de co-autoria a partir de periódicos e eventos científicos, do sistema de consulta, bem como caminhos futuros para o projeto *RedeCI*. Mais especificamente, com relação aos resultados preliminares, busca-se responder à seguinte pergunta: Qual é o perfil do autor dos trabalhos em literatura branca (periódicos) ou cinzenta (eventos) na ciência da informação no Brasil?

Como contribuição, espera-se colaborar com o entendimento da produção da área no Brasil, promover a disseminação dos métodos de ARS e ainda estabelecer as bases para continuação e ampliação do projeto *RedeCI*.

Em relação à organização do texto, na Seção 2 são apresentados os conceitos que embasam a pesquisa, assim como trabalhos relacionados; na Seção 3 são apresentados os meios utilizados na criação da base de dados da *RedeCI* e também os métodos de análise adotados; na Seção 4 é mostrada a análise da colaboração científica na área no Brasil, com base nos artigos analisados; finalmente, na Seção 5 conclui-se que a ARS oferece instrumentos valiosos de análise para a ciência da informação e indicam-se caminhos futuros para o projeto

RedeCI, visando ampliar a base de publicações e os métodos de análise disponíveis.

2. Visão geral da ARS e da análise de co-autoria na *RedeCI*

A *RedeCI*, através da elaboração de uma base de dados de redes de colaboração científica, permite a aplicação de métodos de análise de redes sociais, bem como métodos bibliométricos tradicionais, na análise da produção da área de ciência da informação.

2.1 A ARS e a CI

“Uma rede social (do inglês social network) consiste de um ou mais conjuntos finitos de atores [e eventos] e todas as relações definidas entre eles” (WASSERMAN, FAUST, 1999).

Um *ator* em ARS é uma unidade discreta que pode se apresentar de diferentes formas: como uma pessoa, ou um conjunto discreto de pessoas, agregados em uma unidade social coletiva, como subgrupos, organizações e outras coletividades. Como se pode ver, o conceito de ator é flexível, permitindo diferentes níveis de agregação, o que permite sua adequação a diferentes problemas de pesquisa. Exemplos de atores são: um indivíduo, um pesquisador, um autor, uma empresa e agências de serviço público de uma cidade.

Os *atributos* de um ator são suas características individuais. Embora o foco da ARS sejam os laços relacionais, não há nada que impeça que os atributos sejam analisados, mesmo que com auxílio de outros métodos estatísticos que estão além da ARS. O conjunto de atributos de uma rede é denominado *composição* da rede social.

O *laço relacional (relational tie)* – também denominado simplesmente *laço* ou *ligação (linkage)* – é responsável por estabelecer a ligação entre pares de atores. Uma *relação* em uma rede (*relation*) define todo o conjunto de laços que respeitam o mesmo critério de relacionamento, dado um conjunto de atores. As relações têm duas propriedades importantes que devem ser consideradas nas pesquisas e que condicionam os métodos de análise de dados disponíveis, que são: direcionamento, podendo ser direcionais, caso no qual têm um ator como transmissor e outro como receptor, ou não-direcionais, caso no qual a relação é recíproca; valoração, podendo ser dicotômicas, o que implica em sua presença ou ausência, ou valoradas, com valores discretos ou contínuos. A *distância geodésica*, $d(n_i, n_j)$, entre um par de nós é o número de laços que existe no caminho mais curto entre eles. O *diâmetro* de um grafo é a maior distância geodésica entre todos os pares de nós presentes.

Um grafo não direcionado $G(N, L)$ (*undirected graph*) consiste de dois conjuntos de informações: um conjunto de *nós (node)*, representado por $N = \{n_1, n_2, \dots, n_g\}$ e um ou mais conjunto(s) de *linhas ou arestas (edge)*, representadas por $L = \{l_1, l_2, \dots, l_L\}$ entre pares de nós. O número total de nós existente em um grafo é representado por **g** e o número total de linhas por **L**. A linha (não direcionada) responsável por ligar os nós n_i e n_j é representada por $l_k = (n_i, n_j)$, sendo que $l_k = l_q = (n_i, n_j) = (n_j, n_i)$. Em um grafo (não direcionado), o grau (*nodal degree*), denotado por $d(n_i)$, é o número de linhas incidentes em um nó, ou ainda, de forma equivalente, o número de nós adjacentes a ele. O grau de um nó pode variar de **0**, caso no qual o nó é isolado, até **g – 1**, caso no qual o nó está em contato com todos os demais nós do grafo. No caso de um grafo direcionado,

considera-se o grau de entrada (*indegree*), $d_i(n_i)$, e o grau de saída (*outdegree*), $d_o(n_i)$, de acordo com a direção dos arcos que chegam ou parte do nó.

Pode-se definir algumas medidas de importância de um nó em uma rede, como grau centralidade (*degree centrality*); grau de proximidade (*closeness centrality*) e grau de intermediação (*Betweenness centrality*). Na ARS, essas características estruturais mostram a posição de cada nó. O *grau de centralidade para um ator* é dada por $C_D(n_i) = d(n_i)$; ou seja, é simplesmente o grau do nó. Tal medida pode ser normalizada, a fim de ter um valor entre 0 e 1 e permitir a comparação entre atores de redes diferentes, dividindo-se o grau do nó pelo grau máximo que um nó pode ter, ou seja o número de nós no grafo menos 1 (o próprio

nó), chegando-se a $C'_D(n_i) = \frac{d(n_i)}{g-1}$, sendo $0 \leq C'_D(n_i) \leq 1$. O *grau de proximidade* de

um ator mede o quanto o nó que representa o ator está próximo de todos os demais nós da rede. Para calcular a centralidade de proximidade se soma a distância geodésica do nó em relação a todos os demais nós do grafo e depois inverte-se, uma vez que, quanto maior a distância, menor a proximidade, chegando-se à

seguinte fórmula $C_C(n_i) = [\sum_{j=1}^g d(n_i, n_j)]^{-1}$. A normalização do índice, para o mesmo

ficar na faixa de valores entre 0 e 1, o que pode ser feito multiplicando-se $C_C(n_i)$ por

$g-1$, obtendo-se $C'_C(n_i) = \frac{(g-1)}{[\sum_{j=1}^g d(n_i, n_j)]} = (g-1)C_C(n_i)$. O *grau de intermediação*, por

seu turno, analisa o quanto um nó está no caminho geodésico entre outros nós. Seja g_{jk} o número de caminhos geodésicos (mais curtos e de mesmo tamanho) que ligam os nós j e k , e $g_{jk}(n_i)$ o número de tais caminhos, no total de g_{jk} , que passa

pelo nó n_i . O índice de centralidade de intermediação $C_B(n_i) = \sum_{j < k} \frac{g_{jk}(n_i)}{g_{jk}}$, que mede,

para um nó n_i , a soma de probabilidades de o mesmo estar no caminho geodésico entre todos os demais nós do grafo. Para normalizar tal índice, deve-se dividi-lo pelo seu máximo possível, que é o número de pares de nós no grafo que não incluem n_i , ou seja, $(g-1)(g-2)/2$, chegando-se ao seguinte índice de centralidade de

intermediação normalizado por ator $C'_B(n_i) = \frac{C_B(n_i)}{(g-1)(g-2)/2}$.

A ARS utiliza *softwares* para visualização de grafos a fim de se obter imagens das redes sociais. Geralmente, no caso dos *softwares* disponíveis para ARS, os dados são fornecidos através de matrizes, mas podem ser exibidos através grafos. A visualização gráfica por si só pode oferecer novas informações e *insights* para pesquisadores.

Existe, entre os estudiosos da ARS, uma discussão a respeito de sua fundamentação teórica. Muitas das medidas usadas para descrever a posição dos indivíduos ou atores na rede, como as de centralidade, são objetos de pesquisas que visam trazer elementos para essa fundamentação, assim como criar meios de testar sua validade. Essas medidas visam associar a posição do indivíduo em relação à estrutura da rede e podem ser associadas à influência interpessoal, ao

acesso à informação, às relações de poder, dentre outras. Buscam-se medidas que possam representar a posição do indivíduo na estrutura independente do contexto, isto é, as medidas deveriam oferecer a mesma informação sobre diferentes redes (FRIEDKIN, 1991). No caso das redes de co-autoria, as medidas de centralidade poderiam ser associadas à influência dos autores em relação aos demais.

O uso da ARS vem crescendo significativamente nos últimos 20 anos. Tal crescimento vem ocorrendo em função do aumento da quantidade de dados disponíveis para análise, do desenvolvimento nas áreas de informática e processamento de dados – com o conseqüente aumento do poder computacional à disposição dos pesquisadores –, e da ampliação dos assuntos de interesse e das áreas de conhecimento que utilizam a ARS. Alguns pesquisadores demonstraram essa tendência a partir de pesquisas realizadas em base de dados de artigos científicos e programas de pesquisa. Consultando três bases de dados (*Sociological Abstracts Database*, *Medline Advanced* e *PsycINFO*), Otte e Rousseau (2002) comprovaram, para o período de 1974 a 2000, o crescimento linear do número de artigos publicados anualmente cujo assunto fosse a análise de redes sociais¹. Foram detectados dois momentos nos quais o crescimento no número de artigos se amplia: 1981, graças à publicação de inúmeros manuais sobre o tema e ao desenvolvimento de aplicativos (*softwares*) e 1993, em função da *Web*². O total de artigos sobre o tema, acumulados nas bases pesquisadas, foi multiplicado por 20 entre 1981 e 1999. Além disso, os autores concluíram que o número de assuntos estudados com essa metodologia também se ampliou. No período inicial, a maior parte dos artigos enfocava a família ou a socialização dos indivíduos, enquanto no período final, assuntos tais como sociologia da saúde e da medicina foram encontrados em maior quantidade (OTTE, ROUSSEAU, 2002, p. 446-447). Tomando-se por base a classificação de temas na área de sociologia (*Sociological Abstracts Classification Scheme*), o número de códigos presentes nos artigos passou de cerca de 20 em 1984 para mais de 50 em 1999.

O tema da ARS não é, em termos internacionais, uma proposta nova para a ciência da informação. Para saber se o mesmo fenômeno se repetia na área, Otte e Rousseau (2002) fizeram uma revisão de estudos, aplicando a metodologia de ARS, situando-os a partir do início dos anos 70 e relacionando-os, principalmente, com redes de informação, redes de co-autoria, de pesquisadores e de citações. Posteriormente, combinando a base de dados LISA - *Library and Information Science Abstracts* - e da lista dos principais autores sobre o tema de ARS, pesquisaram sua relevância na ciência da informação. Dos 47 autores mais prolíficos (com 6 ou mais artigos), 12 haviam escrito trabalhos presentes nessa base de dados (independente de serem os primeiros autores).

A pesquisa não alcançou outras áreas do conhecimento que, também, são de interesse da CI. Numa revisão da literatura sobre o paradigma de redes sociais na área de pesquisa organizacional, Borgatti e Foster (2003) comprovam o crescimento da pesquisa em várias áreas do conhecimento. Eles aprofundam a revisão em relação à área de administração e gestão (*Management*), e propõem uma classificação das subáreas de pesquisas em gestão organizacional que utilizam a ARS. Em comum com a área de pesquisa de gestão da informação na área de CI,

¹ Otte e Rousseau utilizaram como filtro para a consulta utilizada na pesquisa o campo *subject heading* das publicações indexadas, buscando o conteúdo '*social network analysis*' (OTTE; ROUSSEAU, 2002, p. 444) e os códigos presentes no *Sociological Abstracts Classification Scheme* (OTTE; ROUSSEAU, 2002, p. 453).

² Embora a *Web* já fosse utilizada como uma rede para conectar pesquisadores desde a década de 1970, é a partir da década de 1990 que a *Web* se implanta. O termo *Web* aqui é uma abreviatura para a "*World Wide Web*".

apontam-se os problemas relacionados com a análise dos fluxos de informação entre os empregados, gerentes, diretores e departamentos dentro de uma organização empresarial.

De uma maneira geral, destacam-se duas grandes linhas de pesquisa que utilizam a ARS e tratam de assuntos relacionados aos problemas de pesquisa na área: i) análise de redes de pesquisadores e de artigos científicos, agregando-se a possibilidade de análise social às técnicas de bibliometria e cientometria; ii) estudos relacionados ao conceito de capital social de pessoas em determinados grupos ou organizações, associados às informações e conhecimentos relevantes para o seu progresso.

2.2 Bibliometria, co-autoria e cooperação

A bibliometria é descrita como um campo de pesquisa verdadeiramente interdisciplinar. Sua metodologia inclui elementos da matemática, ciências sociais e naturais, engenharia, incluindo, ainda, as ciências da vida (GLÄNZEL, 2003). Uma das subáreas de estudo se relaciona com o estudo das co-autorias como indicadores da produção científica. Na verdade, existem várias formas de cooperação entre os cientistas e pesquisadores e essa é apenas uma delas, sendo que a compreensão final da organização social de uma área só poderia ser compreendida com a adição dessas outras formas de relacionamento (CRANE, 1972). No entanto, cada abordagem fornece pistas para as demais e permite testar algumas suposições sobre o comportamento social dos pesquisadores.

Apesar do crescimento acentuado das publicações com múltiplos autores, frutos da crescente cooperação entre instituições e a ampliação das pesquisas multidisciplinares, deve-se levar em conta que existem vários fatores associados à indicação de um colaborador que transcendem a colaboração *per se*. Glänzel (2003), baseando-se nos estudos de Beaver (2001), enumera 18 motivos pelos quais os cientistas cooperam entre si, sendo que alguns estão associados à montagem de redes ou colégios invisíveis, associação com pesquisadores de maior renome ou à obtenção de prestígio. Embora sem entrar na descrição da metodologia de ARS, várias análises são feitas a partir das redes de co-publicação e co-autoria, inclusive entre países. Na descrição de ferramentas bibliométricas, ao final da publicação, observa-se que várias delas combinam as técnicas de bibliometria tradicional com a ARS (GLÄNZEL, 2003).

Na mesma linha, Katz e Martin (1997), consideram que a co-autoria é apenas um indicador parcial da colaboração entre pesquisadores. Inicialmente os autores analisam o significado de colaboração em pesquisa, mostrando que se trata de um termo sem um sentido claro, de uso geral. Eles apontam as dificuldades de se medir a colaboração, mesmo com entrevistas e questionários, em função da complexidade da natureza da interação humana. O artigo analisa ainda a premissa que a colaboração é positiva para os resultados das pesquisas, concluindo que pouco foi feito para explicar como e porque ocorre a colaboração. Com relação ao aumento do número de autores por publicação, e fazendo uma combinação do que vários pesquisadores na área detectaram, são destacadas dez razões.

O aumento da participação de vários pesquisadores de diferentes formações nas pesquisas evidencia a sua característica multidisciplinar e explica o aumento da colaboração, mas outros fatores, como a área de conhecimento, os contatos informais, a distância física e a existência de colégios invisíveis são fatores relevantes. Embora não mencionada claramente, a preocupação com as redes de colaboração aparece na menção aos colégios invisíveis e no compartilhamento de

conhecimento tácito. A distinção entre colaboração e co-autoria poderia ser estudada com auxílio da ARS e a comparação dos elementos e das topologias desses dois tipos de redes.

Vários estudos foram feitos no Brasil na área, com usos de técnicas de bibliometria e cienciometria, para tentar identificar a produção científica de ciência da informação. Esses termos são, muitas vezes, associados a outros, mais recentes, como infometria e *webometria*. Numa discussão sobre a definição, conteúdo e metodologias subjacentes aos termos mencionados, Macias-Chapula (1998) não mencionou o uso de ARS como uma metodologia a ser utilizada na bibliometria. Deve-se mencionar que o objetivo do artigo era discutir o papel e o uso que vem sendo dado à bibliometria e à cienciometria e as dificuldades e suas limitações com relação às comparações internacionais e a disponibilidade de bases de informações adequadas.

Com o objetivo de analisar a produção científica dos docentes dos cursos de pós-graduação em CI registrados até 1999, Población e Noronha, (2002) realizaram um estudo usando essas técnicas. Inicialmente, com base nos atributos dos docentes (área e ano da titulação, linhas de pesquisa, instituição de origem, dentre outros) montaram o quadro sobre a evolução da formação dos professores vinculados ao programa. Em seguida, associaram os professores segundo as linhas de pesquisa identificadas (22 ao todo) e instituições de origem, assim como a sua produção (1108 documentos no período de 1990 a 1999). O total foi dividido em literatura branca (59,8%) e literatura cinzenta (40,2%). Com relação ao número de autores, a maior parte dos documentos (aproximadamente 74%) tinha um único autor e 15% tinham dois autores. As autoras destacaram, entretanto, uma tendência de ampliação do número de trabalhos com dois ou mais autores, como reflexo da constituição de núcleos de pesquisa multidisciplinares, a ampliação da cooperação entre instituições e a maior valorização dos trabalhos em parceria por parte das agências patrocinadoras de pesquisa. Ainda assim, todas as análises sobre a produção e produtividade foram feitas a partir das informações dos atributos e não era objetivo do trabalho identificar quem eram os colaboradores nos grupos de pesquisa e nos documentos já analisados com múltiplos autores.

Em estudo anterior, mas tendo como base uma única publicação – Ciência da Informação, no período de 1990 a 1999 –, foram feitas análises tomando-se por base os atributos dos artigos, tais como o número de autores, os temas, o volume e fascículo no qual foram publicados (MUELLER, PECEGUEIRO, 2001). Com relação à autoria, aproximadamente 78% dos 248 artigos apresentavam um único autor e, daqueles com múltipla autoria (54 artigos), a maior parte apresentava apenas dois autores. Dessa forma, o número total de autores pesquisados chegou a 270, sendo que 83% do total (225) apresentaram um único artigo nas revistas estudadas. Embora relevante para a apresentação de uma metodologia para a análise de publicações, cujos resultados servem como um dos insumos para a compreensão da CI, o estudo poderia ter maior impacto se a base de publicações analisadas fosse ampliada.

2.3 ARS e as redes de co-autoria

Além do trabalho já mencionado de Otte e Rousseau (2002), outros evidenciam a importância da ARS como ferramenta relevante nos estudos de comunidades de prática e na análise bibliométrica, também podem ser citados. Aplicando a ARS em dois departamentos da mesma área (Biologia), mas de duas diferentes universidades suecas, foram analisadas as redes de co-autoria e de

citações diretas e indiretas (MAHLCK, PERSSON, 2000). Os autores testaram basicamente duas hipóteses (sobre a posição dos pesquisadores mais produtivos e sobre a integração de diferentes grupos de pesquisa) tendo por base a estrutura das redes obtidas dos dados oficiais, ou seja, a estrutura dos grupos de pesquisa.

Na mesma linha, tomando por base as redes de co-autoria de grande porte (inclusive com laços internacionais, numa base de 385 autores) Kretschmer (2004) usa a ARS e as demais informações bibliométricas (produção e produtividade) para definir os atributos dos autores e analisar a posição na rede daqueles com as mesmas características ou atributos. O autor testa basicamente três hipóteses: i) se existe uma conexão entre a estrutura dos *clusters* e a produtividade dos cientistas; ii) se existe uma relação entre a distância geodésica e a produtividade; e iii) se a estratificação social é maior quando são pequenas distâncias geodésicas.

Ainda na linha de análise de redes e estrutura social dos colaboradores e, buscando a compreensão sobre a dinâmica dos grupos de pesquisadores, Yoshikane e Kageura (2004) usam a ARS para estudar a estrutura de cooperação entre pesquisadores japoneses de quatro áreas: duas da área de engenharia (engenharia elétrica e processamento de dados) e duas na área de Química (bioquímica e polímeros). Eles constatarem o crescimento relativo do número de artigos produzidos em co-autoria, assim como no número de autores por artigo, com impactos significativos sobre as redes de cooperação.

Ainda no que diz respeito às redes de colaboração entre cientistas, Newman (2001) constrói redes de co-autoria, para o período de 1995 a 1999, a partir de grandes bases de dados americanas (variando de 13.000 artigos e 12.000 autores na área de ciência da computação, a 2 milhões e 1,5 milhões, respectivamente, na área de pesquisa biomédica). Os resultados evidenciam as redes do tipo mundo pequeno (*small-worlds*, (WATTS, 1999)); ou seja, dois cientistas escolhidos aleatoriamente estão separados por um pequeno número de passos (a distância geodésica é curta, cerca de 5 ou 6 passos) e a presença de *clusters* (em todas as bases existe um componente ou subconjunto com cerca de 50 a 80% dos autores, dependendo da área, ou seja, a maioria se conecta entre si através de autores intermediários e dois cientistas têm 30% ou mais de probabilidade de colaborarem entre si, se ambos já colaboraram como um terceiro cientista, isto é, formam uma tríade). (NEWMAN, 2001).

No Brasil, aplicando-se a metodologia de ARS, foi feito um estudo sobre a rede de co-autoria englobando os professores do Programa de Pós-Graduação em Ciência da Informação da UFMG, com o objetivo de identificar: i) se ela é densa, com muitos artigos publicados nessa condição, envolvendo a maior parte dos professores; ii) se existe colaboração entre os professores das diferentes linhas de pesquisa (SILVA *et al.*, 2005). Os resultados estariam manifestados nas publicações totais e, especialmente, nas conjuntas. A base de dados foi construída a partir da base do CNPq – *Lattes* (CNPQ, 2005). Além dos resultados provenientes da aplicação da metodologia da ARS, foram feitas análises estatísticas sobre a produtividade dos autores. Os resultados mostraram que a rede de colaboração é bastante esparsa e muito da produção ocorre entre colaboradores preferenciais. O crescimento da produção está, em grande medida, associado à ampliação da pós-graduação, como demonstra o crescimento da produção em conjunto com alunos do próprio Programa.

Com o objetivo de estudar as redes de colaboração científica, Balancieri *et al.* (2005) fazem uma análise da literatura sobre a cooperação científica situando historicamente os principais pontos tratados por diferentes autores ao longo dos

últimos 50 anos. O artigo enumera vários estudos sobre as razões da cooperação, não entrando nos questionamentos sobre o seu significado e sua importância para a produção e a produtividade dos pesquisadores, embora faça menção a autores que tiveram essa preocupação como, por exemplo, Katz e Martin (1997). Ao longo do período, as análises utilizando a metodologia de ARS são apontadas no final da década de 1990 e início da seguinte como, por exemplo, Newman (2001). Em seguida, os autores analisam a importância das novas tecnologias de informação e comunicação - TIC's - na montagem de bases de dados para estudos das redes de colaboração entre cientistas, tomando como exemplo a Plataforma Lattes. Os autores destacam que existem três sistemas na Plataforma que tratam da análise de redes: a) *Lattes Egressos*, b) *Lattes Colaboradores* e c) *Lattes Redes-GP*. Para estudos das redes de co-autoria, o segundo sistema é o mais indicado. De acordo com os autores pode-se montar a rede de co-autoria de cada pesquisador, com uso da ferramenta *Link Analysis*. A base de dados da Plataforma *Lattes* é, inegavelmente, a maior fonte de informações para a análise da organização das comunidades científicas das diferentes áreas de conhecimento. Existe, ainda, um longo caminho a ser percorrido, especialmente com relação ao uso da base estar atualmente ainda restrita a um grupo de pesquisadores. Não se deve esquecer dos problemas de tratamentos dos dados, conforme mencionados em Silva *et al.* (2005), especialmente a inconsistência entre informações que deveriam ser comuns a dois pesquisadores.

O uso das ferramentas de bibliometria no estudo da *Web* é uma tendência crescente, uma vez que a *Web* e as publicações eletrônicas romperam a barreira da publicação tradicional e da colaboração entre pesquisadores, entre o contato informal e a cooperação (GLÄNZEL, 2003). O estudo de *clusters* na *Web*, da formação de redes do tipo *mundo pequeno* (*small Word*), assim como o mapeamento da estrutura de redes na *Web* são áreas de aplicação das técnicas estatísticas da bibliometria (GLÄNZEL, 2003) e da ARS, chegando-se ao que vem sendo denominado de *webometria*.

3. Metodologia

Adotou-se a mesma metodologia de coleta e análise de dados utilizada no projeto *RedeCI*. Como esta é a primeira publicação referente ao tema, é feita uma apresentação geral da *RedeCI*. Em seguida, descrevem-se as medidas de ARS utilizadas.

3.1 Criação de dados da *RedeCI*

A metodologia para criação da base de dados da *RedeCI* ocorreu por meio de três etapas: inclusão manual de periódicos e eventos; transformação de dados e processamento de consultas, que resultaram em diferentes produtos, conforme esboçado na FIGURA 1. A realização destas atividades culminou na criação do sítio *Web*³.

A primeira etapa refere-se à inclusão dos periódicos e eventos, a partir de seus sítios, em um Banco de dados relacional - BDR - e geração do Esquema relacional da *RedeCI* (ER RCI). Os dados do ER RCI são tratados através de um processo de extração, transformação e carga - ETC - , no qual são realizados filtros para eliminar possíveis erros e dados redundantes, além da identificação de autores idênticos com nomes diferentes. A partir da transformação das tabelas é dada a

³ <http://www.redeci.netic.com.br>

carga no armazém de dados (DW RCI), que serve como fonte de dados básica para as outras aplicações da *RedeCI*.

Após tal tratamento, são gerados três *Data Marts* para as diferentes saídas: i) as tabelas apresentadas no sítio com números da base; ii) arquivos de entrada para o aplicativo UCINET (BORGATTI *et al.*, 2002), com o objetivo de gerar as medidas da rede; iii) arquivos para a visualização da rede em forma de grafo, gerada pelo aplicativo NV2D (BO SHI, 2006).

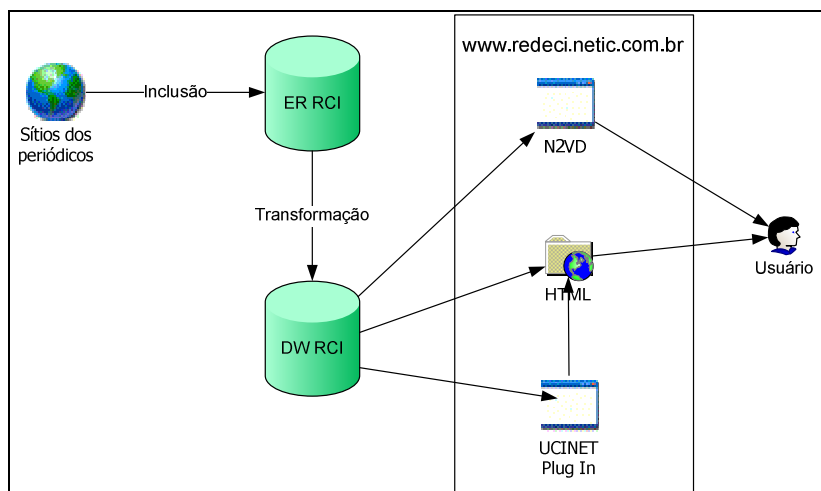


FIGURA 1 - Esquema de funcionamento da RedeCI.

Fonte: Dados da pesquisa

Finalmente, as informações geradas são publicadas no sítio da rede, que conta com a possibilidade de navegação por autores. O sítio apresenta também as últimas mensagens do grupo de discussão, além de bibliografia, *links*, dentre outros.

No contexto da rede, a literatura cinzenta compreende eventos, congressos, colóquios etc. enquanto a literatura branca abrange os periódicos. Autores transientes são aqueles que realizaram somente uma publicação em um determinado grupo de artigos e autores permanentes são aqueles que realizaram mais de uma publicação no mesmo grupo de artigos.

3.2 A ARS e a RedeCI

No caso da RedeCI, o ator é um autor ou pesquisador, sendo que, embora ainda não façam parte da base, seus atributos seriam: as características individuais (idade, sexo) e profissionais (instituição ao qual está vinculada, formação etc.). O laço usado para a montagem da base de dados foi o de co-autoria, sendo eles considerados não direcionais, podendo ser tratados como ora dicotômicos, ora valorados (número de publicações em co-autoria), dependendo do objetivo do estudo. Dessa forma, a RedeCI pode ser representada por um grafo não direcionado, sendo que, por se tratar de uma rede não direcional, as medidas de entrada (*indegree*) e saída (*outdegree*) são idênticas.

Para o presente estudo e, com base nos dados da RedeCI, foram analisadas as seguintes medidas: grau de centralidade (*degree centrality*), grau de proximidade (*closeness centrality*) e grau de intermediação (*betweenness centrality*).

3.3 Outros métodos de análise de produtividade

Além das medidas acima mencionadas, obtidas da ARS, foi feita uma análise sobre a distribuição dos artigos e autores com base na Lei de Lotka. Segundo

Glänzel (2003), Alfred Lotka foi um dos pioneiros da bibliometria ao realizar, em 1926, estudos sobre a produtividade científica na área da química, com base no índice decenal (1907-1916), do *Chemical Abstracts*. Não se trata da única formulação sobre a produtividade científica; ao longo dos anos ela vem sendo criticada ou desafiada por proposições concorrentes (RAVICHANDRA RAO, 1986). Glänzel (2003) considera que a lei de Lotka é, para a descrição da produtividade científica, válida ainda hoje, e a apresenta com a seguinte formulação: “O número de (autores) que fazem n contribuições é cerca de $1/n^2$ daqueles que fazem apenas uma; e a proporção de colaboradores que fizeram apenas uma (contribuição) é de 60%”. (GLÄNZEL 2003, p.6).

Olhando de outra forma, ele mostrou que o número de autores com n publicações em uma bibliografia pode ser descrito como uma lei de potência da forma C/k^β . Ele demonstrou que β tende para 2 e, nesse caso, C seria igual a $6/(\pi)^2$, ou seja, aproximadamente 0,61. Assim, se uma bibliografia pode ser descrita por essa lei de potência, aproximadamente 61% dos autores teriam contribuído com apenas uma publicação. Assim: $f(k) = C/k^\beta, k = 1, 2, \dots$,

Reescrevendo-se a equação como uma distribuição estatística (de forma que a soma de todos os n seja igual a 1), obtém-se: $\sum_{k=1}^{\infty} C/k^\beta = 1$ (ROUSSEAU, ROUSSEAU, 2000).

4. Resultados preliminares

A seguir são apresentados os resultados referentes à primeira versão da base de dados da *RedeCI* e a análise preliminar da produção e da co-autoria.

4.1 Os números da base de dados da *RedeCI*

A base da *RedeCI* contava em dezembro de 2005 com 1309 autores e 1361 artigos, gerando uma proporção de menos de um artigo por autor (0,961). Para a literatura branca, foram considerados periódicos de circulação nacional disponíveis na *Web* e avaliados com classificação *A* pelo *Qualis* em sua classificação relativa aos dados de 2004 (QUALIS, 2005). Para a literatura cinzenta foram considerados os eventos de alcance nacional que levam a palavra *ciência da informação* em seu nome. Uma vez delimitada a literatura, foram incluídos todos os números referentes disponíveis na *Web*. A TABELA 1 apresenta o nome do periódico, o primeiro e último números disponíveis na base, e as respectivas quantidades de artigos e autores.

TABELA 1 - Periódicos e eventos disponíveis na base.

Periódico e eventos	Primeiro Número	Último Número	Quantidade de Artigos	Quantidade de Autores
Ciência da Informação	v. 24, n.1, 1995	v. 34, n. 3, 2005	337	469
CINFORM	n. 4, 2003	n. 6, 2005	127	195
Datagramazero	v. 1, n. 0, 1999	v. 6, n. 6, 2005	160	167
ENANCIB	n. 1, 1994	n. 5, 2003	290	363
EncontrosBibli	n. 1, 1996	n. 19, 2005	72	96
Informação e Sociedade	v. 1, n. 1, 1991	v.15, n. 2, 2005	219	256
Perspectivas em Ciência da Informação	v. 1, n. 1, 1996	v.10, n. 2, 2005	156	195

Fonte: Dados da pesquisa

A partir do estado atual da base, foi possível realizar a análise dos autores transientes e permanentes considerados separadamente pelo tipo de literatura –

cinzenta e branca – ou conjuntamente (TABELA 2). Observa-se que a proporção de autores transientes em relação aos permanentes é maior na literatura branca, ou seja, é mais comum que um autor de um trabalho em um evento publique outro trabalho em outro evento. Os autores permanentes e transientes também são considerados independentemente do tipo de literatura na última linha da TAB. 2.

TABELA 2 - Autores permanentes e transientes.

	Transientes	Permanentes	Proporção
Literatura Branca	615	329	1,869
Literatura Cinzenta	239	178	1,342
Ambas	854	507	1,684

Fonte: Dados da pesquisa

Na TABELA 3 verifica-se o número de autores que publicaram trabalhos em eventos diferentes (e não em edições diferentes do mesmo evento) e revistas diferentes (idem). Observa-se que a quantidade de autores que publicam em diferentes periódicos ou eventos é baixa em relação ao todo. Somente 158 autores (12,1%) publicam artigos em mais de uma revista, enquanto somente 23 publicaram nos dois eventos disponíveis na base. Assim, a maioria absoluta dos autores publica trabalhos sempre no mesmo periódico. Ressalta-se que a base possui somente dois eventos disponíveis (CINFORM e ENANCIB), com suas respectivas edições.

TABELA 3 - Quantidade de autores segundo a diversificação das publicações.

	Apenas 1	Apenas 2	3 ou mais
Literatura Branca	806	113	45
Literatura Cinzenta	512	23	-
Ambas	1082	143	104

Fonte: Dados da pesquisa

4.2 ARS usando a base de dados da *RedeCI*

As medidas de centralidade foram tomadas usando-se o UCINET para 949 autores que apresentaram, pelo menos, uma relação de co-autoria. Não foram feitas análises sobre facções ou grupos isolados entre si (componentes da rede). Os resultados demonstraram que o uso dessa ferramenta requer cuidados especiais e que há necessidade de se aprofundar o conhecimento sobre a estrutura da rede, incluindo o conhecimento sobre a área investigada, no caso, a CI. Portanto, os resultados da ARS sobre a base de dados da *RedeCI*, ainda que iniciais, mostram que as medidas de centralidade precisam ser analisadas com cautela e que sua utilização depende dos objetivos propostos. Por exemplo, tomando-se os valores obtidos para as medidas de centralidade, lembrando-se que o grau de centralidade deve ter sua ordem invertida para a análise pretendida, foi feita uma análise de correlação de *Pearson* para os valores encontrados, o que revelou que as medidas não se correlacionam (TAB. 4).

TABELA 4 - Correlação de Pearson para as medidas de centralidade.

Grau de:	Centralidade	Intermediação	Proximidade (inverso)
Centralidade	1,000		
Intermediação	0,356	1,000	
Proximidade (inverso)	0,178	0,275	1,000

Fonte: Dados da pesquisa

Em outras palavras, o conhecimento da posição dos autores segundo uma determinada medida de centralidade não permite projeções sobre as demais (entre as três aqui usadas).

4.3 Análise de produtividade

Os parâmetros de concentração da produção foram estimados e testados com base no aplicativo desenvolvido por Rousseau e Rousseau (2000) e os resultados mostram uma participação de 72%, maior que a prevista, em relação a autores com apenas uma única contribuição (TAB. 5).

TABELA 5 – Parâmetros da Lei de Potência

Estimativas dos parâmetros da lei de potência (Lei de Lotka)	
Parâmetros	
C	0.7233
beta	24.017

Fonte: Dados da pesquisa

A adequação das estimativas dos parâmetros em uma dada distribuição é testada com sucesso por meio do teste de Kolmogorov-Smirnov, calculado com base nos desvios absolutos entre os valores das funções de distribuição observados e as distribuições teóricas (RAVICHANDRA RAO, 1986). Esse teste indicou que os resultados acima eram, estatisticamente, significativos.

Os valores obtidos chamam a atenção para a possibilidade de que grande parte das contribuições seja de autores que concluíram um trabalho de pesquisa referente à elaboração da sua dissertação.

5. Conclusões

Os resultados da análise da produção indicam a concentração de artigos com autoria única e de autores transientes (com um único artigo) na área e o fato de que os diferentes índices de centralidade têm baixo índice de correlação entre si.

Neste momento, ainda existe um longo trabalho a ser realizado na análise da base de dados de co-autoria com base na ARS, abrindo um amplo espaço para estudos sobre representação, com medidas de fácil compreensão e de uso generalizado, que representem a posição dos autores na estrutura, segundo fatores como influência e importância para a área de CI. Além disso, estudos sobre a estrutura completa da rede, sobre a existência de subgrupos (*clusters* etc.), cliques, redes de cooperação interinstitucionais e colégios invisíveis, dentre outros, permitirão um retrato aprofundado na CI no Brasil. Além disso, estudos de interdisciplinaridade e de epistemologia da área podem ser temas de trabalhos futuros.

Deve-se também destacar que todos os autores dos diversos trabalhos citados, assim como se faz aqui, tomam o cuidado de salientar que a ARS e os resultados estatísticos não eliminam a necessidade de uma análise qualitativa aprofundada. Entender a sistemática de desenvolvimento das conexões, dos mecanismos facilitadores da criação e manutenção de fluxos de informações e de conhecimento são problemas de pesquisa da CI.

Na fase atual, a estrutura da base de dados da *RedeCI* permite o estudo da colaboração entre os pesquisadores na forma de redes de co-autoria, isto é, estudos classificados no primeiro grupo mencionado acima (sistemática do desenvolvimento de conexões). Com a introdução dos atributos dos autores na base de dados, especialmente os dados referentes à instituição de interesse e área de pesquisa, permitirão estudos adicionais, como, por exemplo, das redes de colaboração entre pesquisadores de uma mesma área ou, ainda, das redes de colaboração entre as instituições de pesquisa.

Entre os desenvolvimentos futuros para o projeto *RedeCI* estão previstos a inclusão de atividades como a busca no currículo Lattes (CNPQ, 2005), no Google Scholar (GOOGLE, 2006), na lista da ANCIB (ANCIB, 2006), a inclusão de teses e dissertações defendidas em PPGCs e a utilização de ontologias para descrever as conexões entre os diferentes tipos de informações. O crescente papel da *Web* na comunicação formal e informal entre pesquisadores, as publicações *on-line* e as iniciativas do tipo *Open Archive* indicam a necessidade de uma crescente integração entre a ARS, a bibliometria e as aplicações de *web* semântica, sob pena de não se analisar uma parte relevante da produção e comunicações científica.

Estas informações, disponíveis no armazém de dados, permitirão a geração de consultas e estatísticas mais elaboradas, além da possibilidade de exportar dados em formato estruturado que podem ser utilizados pelos autores em seus respectivos sítios *Web*.

RedeCI: Collaboration and Scientific Production in Information Science in Brazil

This paper presents the methodology and some preliminary results of a coauthorship network, obtained through the use of a technique known like social network analysis - SNA -, in Brazilian Information Science as part of an ambitious project called RedeCI. The numbers of authors per paper are relatively reduced and the numbers of papers with a single author are relevant. The number of authors with a single publication shows results consistent with the literature about scientific productivity, measured by number of papers authored, following Lotka's law. The centrality measures estimated using the social network analysis - SNA - are poorly correlated.

Key-words: Social network analysis; Information science; Bibliometrics; Scientometrics; *Webmetrics*.

Referências

ANCIB. Grupo de discussão da Associação Nacional de Pesquisa em Ciência da Informação. Disponível em <<http://br.groups.yahoo.com/group/ancib/>>. Acesso em 15 fev. 2006.

BALANCIERI, R. *et al.* A análise de redes de colaboração científica sob as novas tecnologias de informação e comunicação: um estudo na Plataforma Lattes. *Ciência da Informação*, Brasília, v. 34, n. 1, p. 64-77, abr. 2005.

BEAVER, D. D. Reflections on scientific collaboration (and its study): past, present, and future. *Scientometrics*, [s.l.], v. 52, n. 3, p. 365-377, Jan. 2001.

BO, Shi. NV2D. Disponível em <http://web.mit.edu/bshi/Public/nv2d/>. Acesso em 12 fev. 2006.

BORGATTI, S. P.; EVERETT, M. G.; FREEMAN, L. C. 2002. *Ucinet for Windows*: Software for Social Network Analysis. Harvard, MA: Analytic Technologies, 2002.

BORGATTI, S. P., FOSTER, P. C. The network paradigm in organizational research: a review and typology. *Journal of Management*, Miami, v. 29, n. 6, p. 991-1013, Dec. 2003.

BROOKES, B. C. The foundations of information science, Part I - Philosophical aspects. *Journal of Information Science*, v. 2, n. 3-4, p.125-133. 1980.

BUSH, V. As we may thing. *Atlantic Monthly*, v. 176, n.1, p. 110-116, July 1945.

CONSELHO NACIONAL DE DESENVOLVIMENTO CIENTÍFICO E TECNOLÓGICO. *Plataforma Lattes*. Brasília, 2005. Disponível em: <<http://lattes.cnpq.br/pl/>>. Acesso em 15 fev. 2006.

CRANE, D. *Invisible Colleges*: diffusion ok knowledge in scientific communities. Chicago: The University of Chicago, 1972.

FLINK: the who is who in the semantic web. Disponível em: <<http://flink.semanticweb.org>>. Acesso em 25 out. 2005.

FRIEDKIN, N. E. Theoretical foundations for centrality measures. *Americam Journal of Sociology*, [s.l.], v. 96, n. 6, p. 1478-1504, May 1991.

GLÄNZEL, W. *Bibliometrics as a research field*: a course on theory and application of bibliometric indicators. [S. I.]: Courses Handout, 2003.

GOOGLE. *Google Scholar*. Disponível em <<http://scholar.google.com>>. Acesso em 15 fev. 2006.

KATZ, J. S.; MARTIN, B. R. What is Research Collaboration? *Research Policy*, v. 26, p. 1-18, 1997.

KRETSCHMER, H. Author productivity and geodesic distance in bibliographic co-authorship networks, and visibility on the Web. *Scientometrics*, [online], v. 60, n. 3,

p. 409-420, Jan. 2004. Disponível em: <<http://www.springerlink.com/app/home/contribution.asp?wasp=463ed3eb2dfb4aa89ed6796b917f7991&referrer=parent&backto=issue,13,22;journal,12,53;linkingpublicationresults,1:101080,1>>. Acesso em 1 abr. 2005.

MACIAS-CHAPULA, C. A. O papel da informetria e da cienciometria e sua perspectiva nacional e internacional. *Ciência da Informação*, Brasília, v. 27, n. 2, p. 134-140, maio/ago. 1998.

MAHLCK, P.; PERSSON, O. Socio-bibliometric mapping of intra-departamental networks. *Scientometrics*, [online], v. 49, n. 1, p. 81-91, 2000. Disponível em <<http://www.springerlink.com/app/home/contribution.asp?wasp=463ed3eb2dfb4aa89ed6796b917f7991&referrer=parent&backto=issue,5,9;journal,47,53;linkingpublicationresults,1:101080,1>>. Acesso em 1 abr. 2005.

MIKA, P. Flink: Semantic Web technology for the extraction and analysis of social networks *Web Semantics: Science, Services and Agents on the World Wide Web*. v. 3, n. 2-3, p. 211-223, Oct. 2005.

MUELLER, S. P. M.; PECEGUEIRO, Claudia Maria. O periódico *Ciência da Informação* na década de 90: um retrato da área refletido em seus artigos. *Ciência da Informação*, Brasília, v. 30, n. 2, p. 47-63, maio/ago. 2001.

NEWMAN, M. E. J. *From the Cover*. The structure of scientific collaboration networks. NATIONAL ACADEMY OF SCIENCES OF USA. *Proceedings...* [online], v. 98, p. 404-409, Jan. 2001. Disponível em: <<http://www.pnas.org/cgi/reprint/98/2/404>>. Acesso em: 01 Mar. 2005.

OTTE, E.; ROUSSEAU, R. Social network analysis: a powerful strategy, also for information sciences. *Journal of Information Science*, Thousand Oaks, v. 28, n. 6, p. 441-453, 2002.

POBLACIÓN, D. A., NORONHA, D. P. Produção das literaturas branca e cinzenta pelos docentes/doutores dos programas de pós-graduação em ciência da informação no Brasil. *Ciência da Informação*, Brasília, v. 31, n. 2, p.98-106, maio 2002.

CAPES. *Qualis*.. Disponível em <<http://qualis.capes.gov.br>>. Acesso em 12 fev. 2005.

RAVICHANDRA RAO, L. K. *Métodos Quantitativos em Biblioteconomia e Ciência da Informação*. Brasília: Associação dos Bibliotecários do Distrito Federal, 1986. 272 p.

REDECI. *Grupo*. Disponível em < <http://br.groups.yahoo.com/group/REDECI> >. Acesso em 12 fev. 2006.

REDECI: Quem é quem na ciência da informação. Disponível em <<http://www.redeci.netic.com.br>>. Acesso em 12 fev. 2006.

ROUSSEAU, B.; ROUSSEAU, R. Lotka: A program to fit a power law distribution to observed frequency data. *Cybermetrics: International Journal of Scientometrics, Informetrics and Bibliometrics*, [s.l.], v. 4, n. 2, p. 1-13, 2000.. Disponível em <<http://www.cindoc.csic.es/cybermetrics/articles/v4i1p4.html>>. Acesso em 25 out. 2005. Aplicativo para cálculo disponível em <<http://www.cindoc.csic.es/cybermetrics/articles/lotka102.exe>>. Acesso em 25 out. 2005.

SARACEVIC, T. Ciência da informação: origem, evolução, relações. *Perspectivas em Ciência da Informação*, Belo Horizonte, v. 1, n. 1, p. 41-62, jan./jun. 1996.

SILVA, A. B. O. *et al.* Estudo da rede de co-autoria e da interdisciplinaridade na produção científica com base nos métodos de análise de redes sociais: avaliação do caso do programa de pós-graduação em ciência da informação - PPGCI / UFMG. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 6., 2005, Florianópolis. *Anais...* Florianópolis [s. n.] 2005.

WASSERMAN, S.; FAUST, K. *Social Network Analysis: methods and applications*. Cambridge: Cambridge University Press, 1999. 857 p. (Structural analysis in the social sciences).

WATTS, D. J. *Small Worlds: the dynamics of networks between order and randomness*. Princeton, New Jersey: Princeton University, 1999.

YOSHIKANE, F.; KAGEURA, K. Comparative analysis of coauthorship networks of different domains: The growth and change of networks. *Scientometrics*, [online], v. 60, n. 3, p. 435-446, jan. 2004. Disponível em: <<http://www.springerlink.com/app/home/contribution.asp?wasp=463ed3eb2dfb4aa89ed6796b917f7991&referrer=parent&backto=issue,15,22;journal,12,53;linkingpublicationresults,1:101080,1;>>. Acesso em 1 abr. 2005.